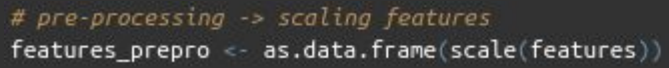
# Data Exploration

Observations over the data:

- All numeric variables. There is no need to convert features
- Different scales. Variables need to be preprocessed so they contribute equally
- To many features. It is necessary to perform feature selection to have predictors that actually contribute to the prediction

The following steps are going to be perform:

- Preprocessing



```
# pre-processing -> scaling features
features_prepro <- as.data.frame(scale(features))
```

- Feature selection
  - Independence, Correlation
  - Subset selection
  - Regularization
- Classifiers training
- Metrics: AUC, MCC

# Feature Selection applied strategies

**Independence, Correlation**

- Removing correlated variables with a correlation bigger that 0.7

**Subset Selection**

- With best subset selection we **perform forwards** and **backwards** method
- For each method the **best adjusted R2**, **CP** and **BIC** are use to **choose** the **predictors**
- **Predictors** from each method are **join** and use for the next step

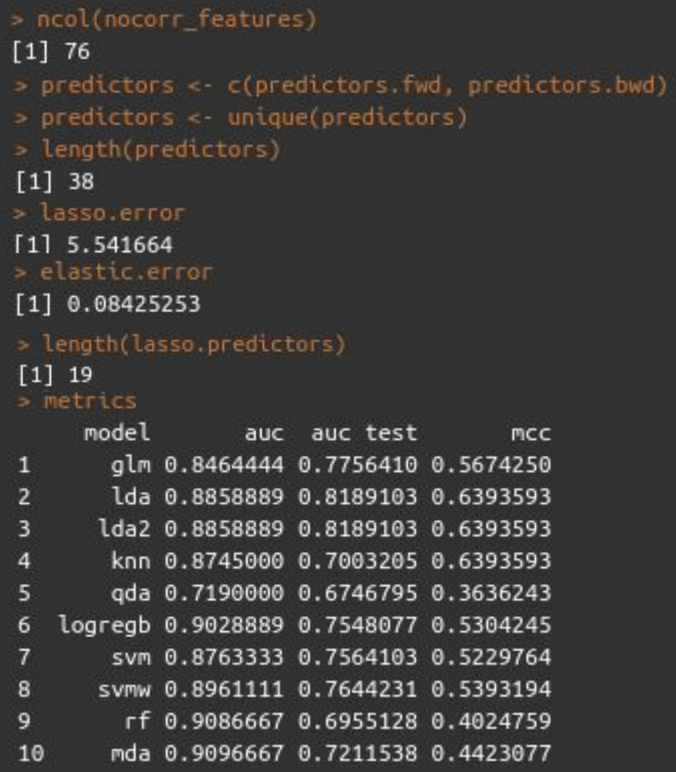
**Regularization**

- **Perform lasso** and **elastic net** in a train subset of the given dataset.
- Compute the **mean square error** for **each** regularization method
- Depending on the **quantity** of **predictors** discard and the **error** one or the other is **use** as the final set of **features**

**Classifiers**

- 10 classifiers are execute with 10 cross validation with AUC as metric measure

# Results Dataset AD vs CTL



```
> ncol(nocorr_features)
[1] 76
> predictors <- c(predictors.fwd, predictors.bwd)
> predictors <- unique(predictors)
> length(predictors)
[1] 38
> lasso.error
[1] 5.541664
> elastic.error
[1] 0.08425253
> length(lasso.predictors)
[1] 19
> metrics
      model       auc  auc test       mcc
1       glm 0.8464444 0.7756410 0.5674250
2       lda 0.8858889 0.8189103 0.6393593
3      lda2 0.8858889 0.8189103 0.6393593
4       knn 0.8745000 0.7003205 0.6393593
5       qda 0.7190000 0.6746795 0.3636243
6   logregb 0.9028889 0.7548077 0.5304245
7       svm 0.8763333 0.7564103 0.5229764
8      svmw 0.8961111 0.7644231 0.5393194
9        rf 0.9086667 0.6955128 0.4024759
10      mda 0.9096667 0.7211538 0.4423077
```

The AUC test and MCC are performed in a validation set extracted from the train dataset given

This to know if the model is overfitting

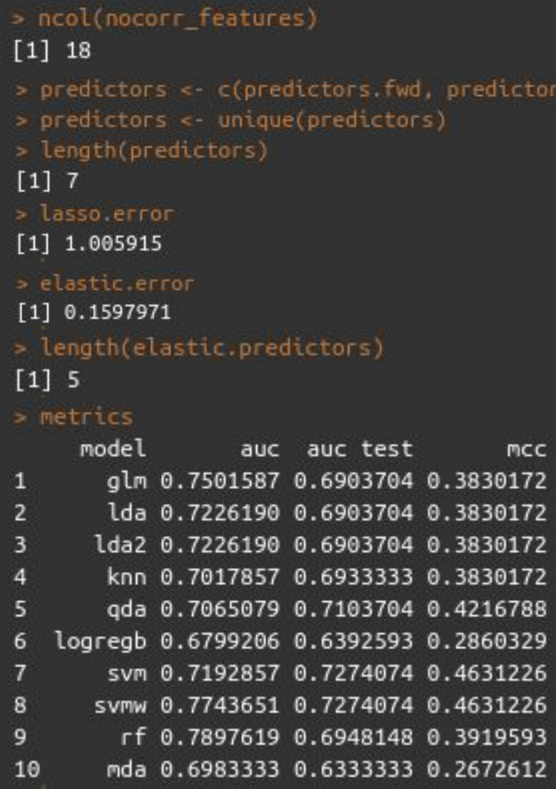
The classifiers used are:

- Generalized logistic regression
- Linear discriminant analysis
- K-nearest neighbor
- Quadratic discriminant analysis
- Logistic regression boost
- Support vector machine
- Support vector machine with weights
- Random Forest
- Mixture and Flexible Discriminant Analysis

The best performing method given the data is:

**Linear discriminant analysis**

# Results Dataset AD vs MCI



```
> ncol(nocorr_features)
[1] 18
> predictors <- c(predictors.fwd, predictors.bwd)
> predictors <- unique(predictors)
> length(predictors)
[1] 7
> lasso.error
[1] 1.005915
> elastic.error
[1] 0.1597971
> length(elastic.predictors)
[1] 5
> metrics
     model       auc  auc test       mcc
1      glm 0.7501587 0.6903704 0.3830172
2      lda 0.7226190 0.6903704 0.3830172
3     lda2 0.7226190 0.6903704 0.3830172
4      knn 0.7017857 0.6933333 0.3830172
5      qda 0.7065079 0.7103704 0.4216788
6  logregb 0.6799206 0.6392593 0.2860329
7      svm 0.7192857 0.7274074 0.4631226
8     svmw 0.7743651 0.7274074 0.4631226
9       rf 0.7897619 0.6948148 0.3919593
10     mda 0.6983333 0.6333333 0.2672612
```

The AUC test and MCC are performed in a validation set extracted from the train dataset given

This to know if the model is overfitting

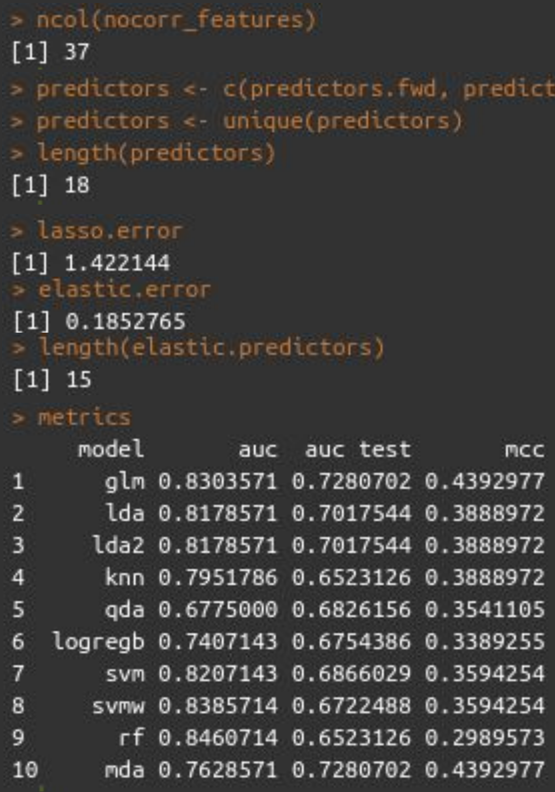
The classifiers used are:

- Generalized logistic regression
- Linear discriminant analysis
- K-nearest neighbor
- Quadratic discriminant analysis
- Logistic regression boost
- Support vector machine
- Support vector machine with weights
- Random Forest
- Mixture and Flexible Discriminant Analysis

The best performing method given the data is:

**Support vector machine**

# Results Dataset MCI vs CTL



```
> ncol(nocorr_features)
[1] 37
> predictors <- c(predictors.fwd, predictors.bwd)
> predictors <- unique(predictors)
> length(predictors)
[1] 18
> lasso.error
[1] 1.422144
> elastic.error
[1] 0.1852765
> length(elastic.predictors)
[1] 15
> metrics
     model       auc  auc test       mcc
1      glm 0.8303571 0.7280702 0.4392977
2      lda 0.8178571 0.7017544 0.3888972
3     lda2 0.8178571 0.7017544 0.3888972
4      knn 0.7951786 0.6523126 0.3888972
5      qda 0.6775000 0.6826156 0.3541105
6  logregb 0.7407143 0.6754386 0.3389255
7      svm 0.8207143 0.6866029 0.3594254
8     svmw 0.8385714 0.6722488 0.3594254
9       rf 0.8460714 0.6523126 0.2989573
10     mda 0.7628571 0.7280702 0.4392977
```

The AUC test and MCC are performed in a validation set extracted from the train dataset given

This to know if the model is overfitting

The classifiers used are:

- Generalized logistic regression
- Linear discriminant analysis
- K-nearest neighbor
- Quadratic discriminant analysis
- Logistic regression boost
- Support vector machine
- Support vector machine with weights
- Random Forest
- Mixture and Flexible Discriminant Analysis

The best performing method given the data is:

**Generalized logistic regression**

# Conclusions

- For data with a lot of features is necessary to perform feature analysis and selection do to the computational limitations and theoretical consequence of using predictors that do not contribute to the model
- Using several models is necessary to been able to compare the fitting of the models
- Realizing overfitting is essential to select the most appropriate model
- Using the metrics that measure what you expect to optimize is a important choice at the moment to define a pipeline